Pause-Focussed Sequential Modelling for Predicting Cognitive Impairment on Limited Data

Benjamin Eyre[1], Aparna Balagopalan[1], Jekaterina Novikova[1]

[1]Winterlight Labs

1. Introduction

Clinical literature suggests that there is a significant difference between the words that healthy and cognitively impaired subjects pause before.

- Pausing in spontaneous speech is indicative of word finding difficulty, which has been shown to be an effective signal for predicting cognitive impairment (CI).
- However, most contemporary methods that attempt to make use of this signal merely quantify the number of pauses a subject makes, or the length of these pauses.
- Is there a way to hone in on the *context* in which pauses occur, and glean more nuanced information about these pauses?

Figure: An example of a pause and its surrounding context from Dementiabank.

We present:

- A novel method of modelling narrative speech as sequences of tokens around pauses, and use this method to classify pauses as being produced by a healthy control (HC) or a subject with cognitive impairment (CI).
- A demonstration that across a variety of sequential machine learning models, the pause-focussed contexts that use sequences of only one word are able to achieve higher accuracy and F1-Score than the same models trained on longer sequences.
- Statistical analyses of features at different token distances from the pause in order to provide an explanation for the differences in model performance.

2. Pause Focussed Modelling

- We use transcripts from Dementiabank, a publicly available dataset of spontaneous picture descriptions from subjects both with and without cognitive impairment.
- For each transcript, we extract each of the pauses, along with the tokens that are one, two, or three tokens away from the pause when possible. A visualization of this is presented in Figure Our final dataset, Utterance, includes every utterance from the Dementiabank transcripts that contained a pause.
- Once these sequences are extracted, a small suite of lexical features are extracted from each of the tokens. Features include length of word, sentiment measures, age of acquisition, etc.

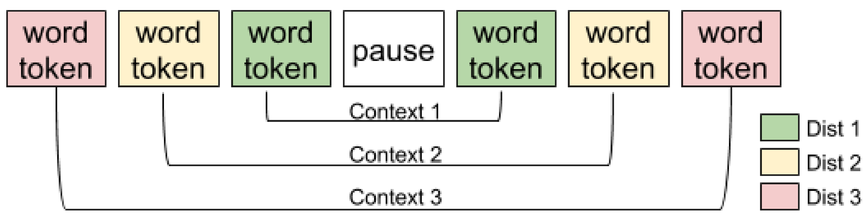

Figure: Visualization of the difference between contexts and distances in a pause-focussed sequence.

Context length	HC	CI	Total
Context 1	333 (33%)	674 (67%)	1007
Context 2	546 (35%)	1003 (65%)	1549
Context 3	559 (35%)	1016(65%)	1575
Utterance	755 (42%)	1060 (58%)	1815

Table: Class distribution of sequences across data subsets in Dementiabank. (https://dementia.talkbank.org)

3. Models

We perform binary classification between HC and CI using three different archetypes of **GRU** based sequential models that produce input for a two layer feed forward neural network that created predictions. The three archetypes are:

- A GRU recurrent neural network, outputting the final hidden state (GRU).
- An GRU with an attention mechanism (ATTEN), where each of the GRU's hidden states are combined into one output context vector.
- A model similar to the previous one, with the exception that the GRU is made bidirectional (Bi-ATTEN).

We experimented with several variations on each of these archetypes, such as using a GRU instead of an LSTM, or by using a weighted loss function. For each model, we perform five-fold cross validation with data Context 1, Context 2, Context 3, and Utterance and compare performance metrics.

For each of the archetypes mentioned, we report the performance for the model with the highest F1-score with respect to the positive class (CI) that also achieved the same specificity as Karlekar *et al*'s CNN-LSTM, one of the best performing models on this task.

4. Results and Reduced Data Set Performance

Model	Context Length	F1	Accuracy	Precision	Specificity	Sensitivity
Bi-ATTEN-L	Context 1	**73.49±0.74**	**61.94±0.64**	69.03±0.29	28.95±1.3	**78.57±0.29**
	Context 2	*67.69±3.22	*58.59±2.41	68.82±0.23	*44.59±5.04	*66.81±6.35
	Context 3	*68.68±2.11	*59.95±1.05	**69.83±1.23**	*46.61±7.69	*67.8±5.45
	Utterance	*64.97±6.01	60.83±1.75	68.82±2.63	***58.63±12.89**	*62.84±12.66
ATTEN-L	Context 1	**72.71±0.76**	**61.36±1.11**	**69.17±0.97**	31.26±3.1	**76.63±0.8**
	Context 2	*70.77±1.33	60.26 ±1.05	*67.81±0.23	*35.5±2.58	*74.03±2.81
	Context 3	*70.95±0.39	60.7±0.37	*68.01±0.22	*36.67±0.63	*74.16±0.72
	Utterance	*66.01±0.87	60.26±0.56	*66.2±0.45	***52.91±1.75**	*65.84±1.79
GRU-S	Context 1	**71.43±1.99**	**60.22±1.37**	**68.88±0.6**	32.19±4.32	**74.26±4.18**
	Context 2	68.96±1.86	59.02±1.77	67.72±1.2	38.32±5.03	70.33±3.91
	Context 3	*68.32±0.76	58.83±0.42	*67.88±0.47	*41.09±2.9	*68.81±2.02
	Utterance	65.07±1.44	59.82±0.67	*66.29±0.73	***54.23±3.28**	*63.98±3.58

Table: Performance metrics (mean±standard deviation) for selected models with highest prediction accuracy that also attained the benchmark specificity. Significant difference when comparing to Context 1 (2-sided t-test) is indicated by an asterisk, and best performance in a metric is denoted by bold.

Models trained on Context 1 consistently outperform the same models trained on other subsets in F1 score and accuracy. The ATTEN-L attains a 1.76% F1 score increase when trained on Context 1 instead of the next best performing data set, Context 3. Similarly, Bi-ATTEN-L achieves a 4.81% F1 score increase when trained on Context 1.

To determine how effective each of the Context lengths are when using a smaller data set, we removed samples from Context 2, Context 3, and Utterance using two different methods:

- For Context 2 and Context 3, remove all sequences that are not also present in Context 1.
- For Context 2, Context 3, and Utterance we deleted random samples until only 1007 samples remained, while maintaining class ratios for each data set.

Model	Context Length	F1-Score	Accuracy	Precision	Specificity	Sensitivity
Bi-ATTEN-L	Context 1	**73.49±0.74**	**61.94±0.64**	69.03±0.29	28.95±1.3	**78.57±0.29**
	Context 2	*69.57±1.96	*59.7±1.21	*70.43±0.77	*41.87±5.67	*68.87±4.44
	Context 3	*71.18±1.46	61.22±1.26	***71.08±1.18**	*41.34±5.4	*71.37±3.31
	Utterance	*54.58±12.88	55.26±5.62	*67.6±3.01	***65.81±16.77**	*48.37±20.35
ATTEN-L	Context 1	72.71±0.76	61.36±1.11	69.17±0.97	31.26±3.1	**76.63±0.8**
	Context 2	72.71±0.63	62.06±0.36	**70.36±0.66**	35.79±3.55	75.26±1.88
	Context 3	**73.35±0.94**	**62.41±1.03**	70.22±1.04	34.28±4.42	76.81±2.14
	Utterance	*65.68±2.31	59.73±2.82	*65.61±2.53	***51.67±4.44**	*65.76±2.29

Table: Average performance metrics for models after each dataset was modified with the first described reduction method.

Model	Context Length	F1-Score	Accuracy	Precision	Specificity	Sensitivity
Bi-ATTEN-L	Context 1	**73.49±0.74**	**61.94±0.64**	**69.03±0.29**	28.95±1.3	**78.57±0.29**
	Context 2	*68.71±3.13	59.18±2.87	68.4±1.2	*41.47±3.07	*69.13±5.28
	Context 3	*67.03±2.93	*58.38±1.96	68.5±0.7	***45.2±5.78**	*65.8±5.99
ATTEN-L	Context 1	**72.71±0.76**	**61.36±1.11**	**69.17±0.97**	31.26±3.1	**76.63±0.8**
	Context 2	*70.28±1.65	59.55±1.96	67.26±1.82	**33.98±5.86**	73.65±2.87
	Context 3	70.99±0.5	60.19±1.1	67.14±1.0	33.07±3.22	*75.32±0.54

Table: Average performance metrics for models after each dataset was modified with the second described reduction method.

The performance in the second table suggests that the samples within Context 1 present a signal differentiating between HC and CI that is stronger with less samples than Context 2, Context 3, and Utterance. The performance in the first table suggests that when using these samples, models are able to exploit this signal while still profiting from the additional tokens added by extending the context. We touch upon this further in the following section.

5. Effect of Context

We performed two sided t-tests for the features of tokens that occupy the same position in reference to the pause. We report an aggregated number of significant features for distance 1, distance 2, and distance 3 from the pause, as demonstrated in the figure in section 2.

Distance	Sign.	# Features
Dist 1	$5.3e^{-71}$	13
Dist 2	$6.5e^{-28}$	12
Dist 3	$1.0e^{-6}$	2

Table: Average of p-values among two most significant features after Bonferoni correction for each distance, along with the total number of significant features.

Distance 1 contained a larger number of significant features with lower p-values, implying that they are more significantly different between the classes. This concentration of significant features is what makes models trained on Context 1 more effective.

6. Conclusions

- Smaller, specific token centred sequences can provide equal or greater CI classification performance than longer sequences, including the entire utterance.
- This strategy can be used to enhance performance on small, noisy speech corpora.
- Enhanced performance on the inner context can be attributed to a smaller amount of noise and higher concentration of significant features at distance 1.

7. References

[1] Pistono *et al.*, Pauses during autobiographical discourse reflect episodic memory processes in early Alzheimer's disease. Journal of Alzheimer's Disease, 50(3):687–698, 2016.
[2] Becker *et al.*, The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. Archives of Neurology ,51(6):585-594, 1994.